LOW RANK ADAPTIVE FRICTION: MEMORY EFFICIENT OPTIMIZATION WITHOUT LINEAR DAMPING

Anonymous authors
Paper under double-blind review

ABSTRACT

iKFAD is a recently proposed optimiser that replaces adaptive learning rates with adaptive friction in the momentum dynamics, yet performs as well as Adam. Its limitation is that the full friction tensor $\xi \in \mathbb{R}^{m\times n}$ carries the same $\mathcal{O}(mn)$ memory overhead per layer as Adam's second-moment buffer. Here we replace iKFAD's friction tensor ξ with a rank-1 outer-product factorisation built from row and column momentum statistics, resulting in Rank-1 iKFAD (R-iKFAD). This reduces the friction memory footprint from $\mathcal{O}(mn)$ to $\mathcal{O}(m+n)$ per layer, which approximately halves iKFAD's total optimiser state. Despite this reduction, R-iKFAD maintains parity in performance with iKFAD: experiments on GPT2-Nano, TinyViT, DistilBERT and GPT2-S confirm that it matches or exceeds iKFAD while nearly halving the memory footprint and remaining comparably robust to hyperparameters. We analyse the continuous-time dynamics in two damping regimes. For linear damping ($\gamma > 0$) we prove exponential convergence under strong convexity. For $\gamma = 0$, the preferred option in our experiments, the friction is generated entirely from past momentum and switches off as the momentum vanishes, so geometric convergence cannot be shown. We nonetheless prove convergence to the minimiser, together with matching upper and lower bounds on the energy: of order t^{-1} when the regularisation scale ϵ_{stab} is zero, and of order $t^{-1/2}$ when it is positive. To our knowledge this is the first convergence rate for a rank-1 factored optimiser in continuous time, and the first such result that does not require positive damping.

1 INTRODUCTION

Adam (Kingma & Ba, 2015) remains the workhorse optimizer of modern deep learning, combining momentum with coordinate-wise second-moment scaling to accelerate convergence on complex loss landscapes. Adam evolves two auxilliary parameters: the first and second moment estimates. Assuming the model has N parameters, the auxilliary states occupy $2N$ total memory cost beyond the N parameters themselves. At large scale, this $2N$ overhead presents a significant hardware bottleneck for large-model training. To mitigate this, Adafactor (Shazeer & Stern, 2018) introduces a rank-1 factorisation of the second-moment matrix. For a matrix weight $W \in \mathbb{R}^{m\times n}$, Adafactor replaces Adam's second-moment matrix $V \in \mathbb{R}^{m\times n}$ with a factored approximation $\hat{V} = RC^T/(\mathbf{1}_m^T R)$, where $R \in \mathbb{R}^m$ and $C \in \mathbb{R}^n$ track the row-wise and column-wise squared gradients, respectively. By eliminating the first-moment buffer, Adafactor reduces the per-layer auxiliary state complexity from $\mathcal{O}(mn)$ to $\mathcal{O}(m+n)$ (achieving memory savings of over 99% when large weight matrices are involved) while maintaining competitive transformer pre-training performance.

A useful approach to analyzing optimizer dynamics is based on continuous-time ordinary differential equations (ODEs) (Da Silva & Gazeau, 2020; Karoni, 2024), where updates are defined as discrete time-stepping approximations. For parameter state vector x, momentum vector p, and second-moment vector ζ, continuous-time Adam is given in Eqs. (1)-(3). While Adam incorporates parameter-wise adaptation into the *position* equation ($\dot{x}$) via $1/\sqrt{\zeta}$, Karoni et al. (2026) demonstrated that coordinate-wise adaptation can alternatively be integrated into the *momentum* equation ($\dot{p}$). This is achieved by replacing the constant damping scalar factor γ with a dynamic, coordinate-wise friction state ξ, resulting in the Individual Kinetic Friction-Adaptive Descent (iKFAD) dynamics (4)-(6)

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

$$\dot{x} = \frac{p}{\sqrt{\zeta} + \epsilon}, \tag{1}$$

$$\dot{p} = -\nabla f(x) - \gamma p, \tag{2}$$

$$\dot{\zeta} = [\nabla f(x)]^2 - \alpha\zeta. \tag{3}$$

$$\dot{x} = p, \tag{4}$$

$$\dot{p} = -\nabla f(x) - (\gamma + \xi) \odot p, \tag{5}$$

$$\dot{\xi} = \frac{[p]^2}{\mu} - \alpha\,\xi. \tag{6}$$

These dynamics require storage of a full adaptive friction tensor ξ alongside the momentum p. They incur an auxiliary footprint of $2N$ scalars, identical to Adam's auxiliary storage. Originally introduced for scalar friction (Karoni et al., 2023),iKFAD allows friction-based adaptation to individual parameters while matching Adam's performance on standard vision and NLP benchmarks. This suggests an apparent gap: Adafactor obtains its saving by factoring the object Adam stores in the position equation, yet no available counterpart exists for the object that iKFAD stores in the momentum equation. Figure 1 arranges the four methods discussed in this paper. The lower-right cell is the new method presented here: Rank-1 iKFAD (R-iKFAD) applies Adafactor's factorisation to the friction tensor, reducing iKFAD's auxiliary state from $2N$ to $N + \sum_\ell (m_\ell + n_\ell)$ scalars while retaining full momentum.

2 Methodology: Rank-1 iKFAD

Adafactor compresses Adam's second moment while Rank-1 iKFAD (R-iKFAD) compresses iKFAD's friction. Writing both as continuous-time systems for a matrix weight $X \in \mathbb{R}^{m\times n}$ clarifies this correspondence, in the same sense that the systems above approximate Adam (Da Silva & Gazeau, 2020) and iKFAD (Karoni, 2024). Adafactor implementations generally omit momentum. We retain this momentum P and refer to the resulting scheme as **Adafactor-m** (left). (Retaining momentum is deliberate as Adafactor is known to suffer training instabilities in certain transformer regimes when momentum is discarded entirely (Shazeer & Stern, 2018).) Both continuous systems use a constant second-moment linear decay coefficient α to be directly comparable. The discrete Adafactor-m baseline follows the schedule $\beta_{2,t} = 1 - t^{-0.8}$ (Appendix B). R-iKFAD (right) maintains the same momentum component and factors the friction matrix instead:

$$\dot{X} = \frac{P}{\sqrt{\hat{\zeta}} + \epsilon}, \tag{7}$$

$$\dot{P} = -\nabla f(X) - \gamma P, \tag{8}$$

$$\dot{R} = [\nabla f(X)]^2\,\mathbf{1}_n - \alpha R, \tag{9}$$

$$\dot{C} = ([\nabla f(X)]^2)^\top \mathbf{1}_m - \alpha C. \tag{10}$$

$$\dot{X} = P, \tag{11}$$

$$\dot{P} = -\nabla f(X) - \tilde{\xi} \odot P - \gamma P, \tag{12}$$

$$\dot{R} = \frac{1}{\mu}[P]^2\mathbf{1}_n - \alpha R, \tag{13}$$

$$\dot{C} = \frac{1}{\mu}([P]^2)^\top \mathbf{1}_m - \alpha C. \tag{14}$$

Here $\odot$ is the element-wise product, $[P]^2_{ij} = P^2_{ij}$, and the divisions in the position equations are element-wise. Both systems carry the same linear damping γ. Each builds a rank-1 object from its factors,

$$\hat{\zeta} = \frac{RC^\top}{\mathbf{1}_m^\top R}, \qquad \tilde{\xi}_{ij} = \frac{R_i C_j}{\mathbf{1}_m^\top R + \epsilon_{\text{stab}}}, \tag{15}$$

with $\epsilon, \epsilon_{\text{stab}} > 0$ small regularization coefficients, the latter ensuring numerical stability near $R = 0$.

Adafactor-m is thus the Adam system with the full second moment ζ replaced by the rank-1 estimate $\hat{\zeta}$, and R-iKFAD is the iKFAD system with the full friction tensor ξ replaced by the rank-1 estimate $\tilde{\xi}$. This is illustrated more concisely in Figure 1. The two methods differ in the same way Adam differs from iKFAD: Adafactor-m's factors aggregate the squared *gradients* and act in the position equation, whereas R-iKFAD's aggregate the squared *momenta* and act as damping in the momentum equation. In both cases the adapted state per layer (the second moment or the friction) falls from $\mathcal{O}(mn)$ to $\mathcal{O}(m+n)$, while full momentum evolution is retained. We emphasize that R-iKFAD is designed to maintain parity with iKFAD (which matches Adam (Karoni et al., 2026) in performance), rather than outperforming it in terms of convergence.

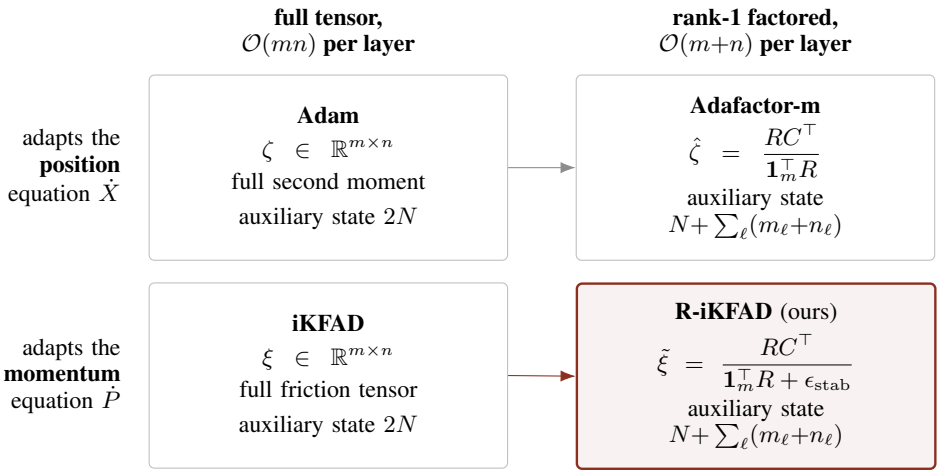

Figure 1: The same rank-1 compression applied in two different equations. Adafactor-m factorizes Adam's second moment ζ which scales the position update while R-iKFAD factorizes iKFAD's friction ξ which damps the momentum update. Both reduce the adapted state per layer from $\mathcal{O}(mn)$ to $\mathcal{O}(m+n)$ while keeping full momentum, so total auxiliary memory falls from $2N$ to roughly N.

Splitting Discretization. To construct a stable discrete-time algorithm, we turn to principles of geometric numerical integration of Hamiltonian systems (Leimkuhler & Reich, 2005), which have proved a robust foundation for discretization of dynamics-based optimisation schemes. Following Karoni et al. (2026), we decompose the vector field of (11)–(14) into four components, three of these being directly integrable. The fourth is easily approximated by an additional splitting step. Specifically we write:

$$\begin{pmatrix}\dot{X}\\ \dot{P}\\ \dot{R}\\ \dot{C}\end{pmatrix} = \underbrace{\begin{pmatrix}P\\0\\0\\0\end{pmatrix}}_{\text{A}} + \underbrace{\begin{pmatrix}0\\-\nabla f(X)\\0\\0\end{pmatrix}}_{\text{B}} + \underbrace{\begin{pmatrix}0\\-\frac{RC^\top}{\mathbf{1}_m^\top R+\epsilon_{\text{stab}}}\odot P\\ \frac{1}{\mu}[P]^2\mathbf{1}_n-\alpha R\\ \frac{1}{\mu}([P]^2)^\top\mathbf{1}_m-\alpha C\end{pmatrix}}_{\text{C}} + \underbrace{\begin{pmatrix}0\\-\gamma P\\0\\0\end{pmatrix}}_{\text{D}}. \tag{16}$$

Sub-flows A, B, and D can be integrated analytically: A gives position drift, B applies the gradient kick, D applies linear friction (becoming the identity when $\gamma = 0$, i.e., R-iKFAD0). For step C, we damp the momentum for half a step using the current factors, update the factors for a full step with momentum held fixed, then damp the momentum for another half step using the updated factors. The composition applied in practice is a first-order Lie–Trotter splitting:

$$\Phi_h = \Phi_h^{\text{D}} \circ \Phi_h^{\text{C}} \circ \Phi_h^{\text{A}} \circ \Phi_h^{\text{B}},$$

where sub-steps are executed in the sequence B $\to$ A $\to$ C $\to$ D (BACD) (Karoni et al., 2026). Position updates use the gradient-kicked momentum $P^{(B)}$ rather than P_n. Table 3 in Appendix A gives the explicit analytical update rules for each sub-step across the complete BACD sequence over time step h, together with the treatment of the coupling inside sub-flow C.

Memory Complexity. For a network whose ℓth parameter matrix is $W_\ell \in \mathbb{R}^{m_\ell\times n_\ell}$, let $N = \sum_\ell m_\ell n_\ell$ denote the total number of scalar model parameters. Figure 1 breaks down the auxiliary optimizer memory footprint required on top of the model weights. For non-2D parameters (biases, LayerNorm scales) that are not rank-1 factorizable, R-iKFAD reduces to element-wise iKFAD, contributing a negligible additional scalar per parameter. R-iKFAD requires approximately N auxiliary scalars in total, effectively halving the optimizer auxiliary memory footprint relative to both iKFAD and Adam. Adafactor-m has a similar auxiliary memory requirement $\approx N$ and is therefore the primary baseline in our experiments.

3 THEORETICAL RESULTS

We next discuss our analysis of the continuous-time system (11)–(14) in both damping regimes. This section states the results and reports the numerical verification. The proofs are in Appendix G.

For $\gamma > 0$, a Lyapunov argument gives exponential convergence under strong convexity.

Theorem 1 (Exponential convergence for $\gamma > 0$)**.** *Suppose that f is m_f-strongly convex and has an L_f-Lipschitz gradient. Let $\gamma, \alpha, \mu, \epsilon_{\mathrm{stab}} > 0$, and let $R(0)$ and $C(0)$ be componentwise nonnegative. Then there are constants $M, \rho > 0$ such that*

$$f(X(t)) - f(X^*) + \|X(t) - X^*\|_F^2 + \|P(t)\|_F^2 + \|R(t)\|_2^2 + \|C(t)\|_2^2 \le Me^{-\rho t} \tag{17}$$

for every $t \ge 0$.

In our experiments the $\gamma = 0$ and $\gamma > 0$ variants performed very similarly, and R-iKFAD0 is preferable since it has one fewer hyperparameter to optimize, which makes the sweep easier. With $\gamma = 0$ the overall adaptive friction remains non-negative but can become small with the momentum. Proving convergence in this setting is challenging, but we have succeeded in doing so under strong convexity.

Theorem 2 (Convergence when $\gamma = 0$)**.** *Suppose that f is strongly convex and has a locally Lipschitz gradient. Let $\alpha, \mu, \epsilon_{\mathrm{stab}} > 0$, and suppose that $R(0)$ and $C(0)$ are componentwise nonnegative. Then every solution of* (11)–(14) *with $\gamma = 0$ converges to*

$$(X, P, R, C) = (X_\star, 0, 0, 0),$$

where $X_\star$ is the unique minimizer of f.

The decay at $\gamma = 0$ is algebraic rather than exponential, meaning that the error falls like t to some power. The power depends on whether or not a stabilizing parameter ϵ_{stab} is included. Both this theorem and the numerical evidence that follows are stated in terms of the energy

$$\mathcal{E}(t) = f(X(t)) - f(X^*) + \tfrac{1}{2}\|P(t)\|_F^2, \tag{18}$$

the sum of the objective gap and the kinetic energy term, which is nonnegative and vanishes exactly at $(X^*, 0)$.

Theorem 3 (Algebraic rates)**.** *Suppose f is strongly convex, its gradient is locally Lipschitz, and it is C^2 in a neighbourhood of X^*. Let $\gamma = 0$ and $\alpha, \mu > 0$, and let the initial factors be component-wise nonnegative.*

If $\epsilon_{\mathrm{stab}} > 0$ and $\mathcal{E}(0) > 0$, there are $a, A > 0$ such that

$$\frac{a}{\sqrt{1+t}} \le \mathcal{E}(t) \le \frac{A}{\sqrt{1+t}}, \qquad t \ge 0.$$

If $\epsilon_{\mathrm{stab}} = 0$, assume in addition that $R(0) = C(0) = 0$, and use the convention $\tilde{\xi} = 0$ at $\mathbf{1}_m^\top R = 0$. If $\mathcal{E}(0) > 0$, there are $b, B > 0$ such that

$$\frac{b}{1+t} \le \mathcal{E}(t) \le \frac{B}{1+t}, \qquad t \ge 0.$$

The two regimes are separated by a crossover when $\mathcal{E}$ is of order $\alpha\mu\epsilon_{\mathrm{stab}}$. Section 3.1 gives a formal averaging calculation that identifies the exponents, while Appendix G.3.1 proves those results, and Section 3.2 verifies them numerically. The two-sided bounds establish $t^{-1/2}$ as the asymptotic decay rate for every $\epsilon_{\mathrm{stab}} > 0$, motivating a small stabiliser to delay the crossover to this slower regime.

3.1 FORMAL RATE CALCULATION

We use a formal averaging calculation to estimate the decay. The exponents it produces are rigorously established in Appendix G.3.1. Suppose that P oscillates faster than the factors change. For $\epsilon_{\mathrm{stab}} = 0$, start with $\sum_k R_k > 0$. Replacing the oscillatory terms by the corresponding averages gives

$$\dot{\mathcal{E}} \simeq -\frac{\mathcal{E}^2}{\alpha\mu}.$$

The reduced equation admits the solution

$$\mathcal{E}(t) \simeq \frac{\alpha\mu}{t}.$$

When $\epsilon_{\text{stab}} > 0$ and the factors are small relative to ϵ_{stab}, the same calculation gives

$$\dot{\mathcal{E}} \simeq -\frac{\mathcal{E}^3}{\alpha^2\mu^2\epsilon_{\text{stab}}}, \qquad \mathcal{E}(t) \simeq \alpha\mu\sqrt{\frac{\epsilon_{\text{stab}}}{2t}}.$$

The two reduced regimes meet when $\mathcal{E}$ is of order $\alpha\mu\epsilon_{\text{stab}}$. This calculation relies on time-scale separation and an averaging closure. Neither assumption is needed for Theorem 3, which obtains the same exponents without them.

3.2 Numerical validation of proved results

We tested the dynamics of $\gamma = 0$ in a strongly convex quadratic, $f(X) = \frac{1}{2}X^\top AX$ with $A = BB^\top/d + 0.8I$ and $d = mn$, where B has independent standard normal entries, so the smallest eigenvalue of A is at least 0.8. We monitor the energy $\mathcal{E}(t) = \frac{1}{2}X^\top AX + \frac{1}{2}\|P\|_F^2$. The decay for both stabilization settings may be found in Figure 2. Table 1 gives the slopes of $\log\mathcal{E}$ against $\log t$, obtained by ordinary least squares on successive decades out to $T = 10^7$ for three different matrix sizes. Both of the powers specified by our theorem are recovered and neither apparently depends on the matrix dimension. The approach to slope $-1/2$ is slower for larger mn, and the crossover point illustrates the reason for this: the second regime begins only once $\mathcal{E} \lesssim \alpha\mu\,\epsilon_{\text{stab}}$, and the constant in the first-regime law grows with mn, so the crossing occurs later for larger systems. The prefactors scale with α, μ and ϵ_{stab} just as the calculation of Section 3.1 predicts. Appendix E reports the measured ratios and Appendix F gives the integrators, tolerances, step-sizes, seeds and initialization.

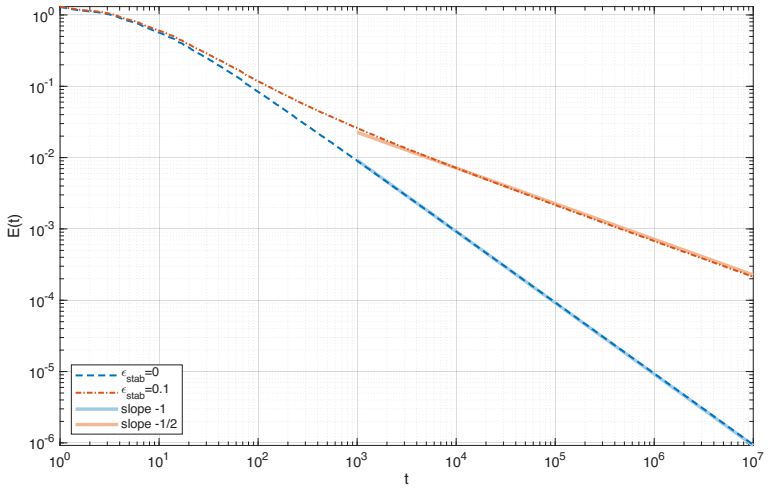

Figure 2: Energy decay from BACD splitting with $m = 4$, $n = 3$, $\gamma = 0$, $\alpha = \mu = 1$, $h = 0.02$, and final time 10^7. Both runs have $R(0) = C(0) = 0$. The blue dashed curve has $\epsilon_{\text{stab}} = 0$, and the orange dash-dotted curve has $\epsilon_{\text{stab}} = 0.1$. The solid lines have slopes -1 and $-1/2$, respectively.

3.3 Related Work: Algebraic Rates and Factored Optimisers

Continuous-time studies with damping such as Su et al. (2016) and Attouch et al. (2018) prescribed as a function of time have yielded algebraic rates. Our results are proven for a different setting given that our damping is determined by evolving factors and can vanish with them. Our damping model is also nonlinear and thus more complicated. In terms of factored schemes, Adafactor (Shazeer & Stern, 2018) introduced rank-1 factorization of optimizer state. Hong & Lin (2025) analyzed a discrete Adafactor method for non-convex objectives. Nguyen et al. (2025) introduced H-Fac,

Table 1: Local slopes of $\log \mathcal{E}$ against $\log t$ on successive decades, from the BACD splitting at $h = 0.02$ with $\alpha = \mu = 1$, integrated to $T = 10^7$. Both exponents are independent of the matrix dimensions.

	$[10^3, 10^4]$	$[10^4, 10^5]$	$[10^5, 10^6]$	$[10^6, 10^7]$
$\epsilon_{\mathrm{stab}} = 0$ (proved exponent -1)				
$m = n = 2$	-0.955	-0.980	-0.986	-0.989
$m = 4, n = 3$	-0.991	-0.996	-0.996	-0.998
$m = 6, n = 5$	-0.993	-0.997	-0.998	-0.999
$\epsilon_{\mathrm{stab}} = 0.1$ (proved exponent $-1/2$)				
$m = n = 2$	-0.499	-0.497	-0.496	-0.497
$m = 4, n = 3$	-0.558	-0.518	-0.505	-0.501
$m = 6, n = 5$	-0.593	-0.530	-0.509	-0.503

which factorizes both the momentum and scaling estimators, and prove asymptotic convergence for its continuous-time dynamics with positive damping. These results do not come near providing rates for $\gamma = 0$ situation. KFAD (Karoni et al., 2023) and iKFAD (Karoni et al., 2026) were analyzed only in the linearly damped case $\gamma > 0$. Our $\gamma = 0$ results carry over to both. In each, the friction obeys the same non-negative linear filter as the factors R_i and C_j: a single scalar driven by $\|P\|_F^2$ in KFAD, and one entry ξ_{ij} driven by P_{ij}^2 for each parameter in iKFAD. Neither method has a stabilizer, so only the t^{-1} rate applies.

4 Numerical experiments

We set $\gamma = 0$ for both iKFAD and R-iKFAD throughout (denoted **iKFAD0** and **R-iKFAD0** respectively), following Karoni et al. (2026). In the ablation of Section 4.2, adding $\gamma > 0$ gave no consistent improvement under our tested procedure. We compare both against Adam and Adafactor-m across four benchmarks spanning problems in image classification and language modelling at scales from 0.8M to 124M parameters:

- **GPT2-Nano (0.8M)** (Karpathy, 2022) on Shakespeare character-level language modelling.
- **TinyViT (4M)** (Wu et al., 2022) on CIFAR-10 (Krizhevsky et al., 2009) image classification.
- **DistilBERT (67M)** (Sanh et al., 2020) on SST-2 (Socher et al., 2013) sentiment classification (pretraining from scratch).
- **GPT2-S (124M)** (Radford et al., 2019) on OpenWebText pretraining (nanoGPT architecture) (Karpathy, 2022).

To focus on the optimization dynamics, we use a constant learning rate, we disable weight decay, and we omit additional regularization throughout. We use Adafactor-m rather than the conventional momentum-free configuration because its auxiliary-memory requirement is similar to that of R-iKFAD. All hyperparameters are tuned via Optuna (Akiba et al., 2019) (TPE sampler) with equal sweep budgets of approximately 80 trials per optimiser per model. GPT2-Nano and TinyViT received substantially more trials. The stabilizer is fixed at $\epsilon_{\mathrm{stab}} = 10^{-16}$ in every run and is never tuned. See Appendix B for full configurations. As with iKFAD, the optimal scale of μ varies by several orders of magnitude between tasks due to differing momentum scales. We therefore recommend a wide-range initial sweep over μ. Note that Adafactor-m requires tuning only two hyperparameters, whereas Adam, iKFAD0, and R-iKFAD0 all require three. The fixed trial budget hence favours Adafactor-m's results disproportionately. All benchmarks report mean $\pm$ standard deviation over 10 seeds, except GPT2-S, which uses a single seed due to computational constraints. The optimizer auxiliary state is the memory an optimizer adds on top of the model parameters. Table 5 in Appendix C reports the absolute values on the four benchmarks.

4.1 Main Results

Figure 3 shows training and test loss across all benchmarks, and Figure 7 in Appendix D the corresponding accuracies. Table 2 reports the test loss of the model saved at the lowest validation loss.

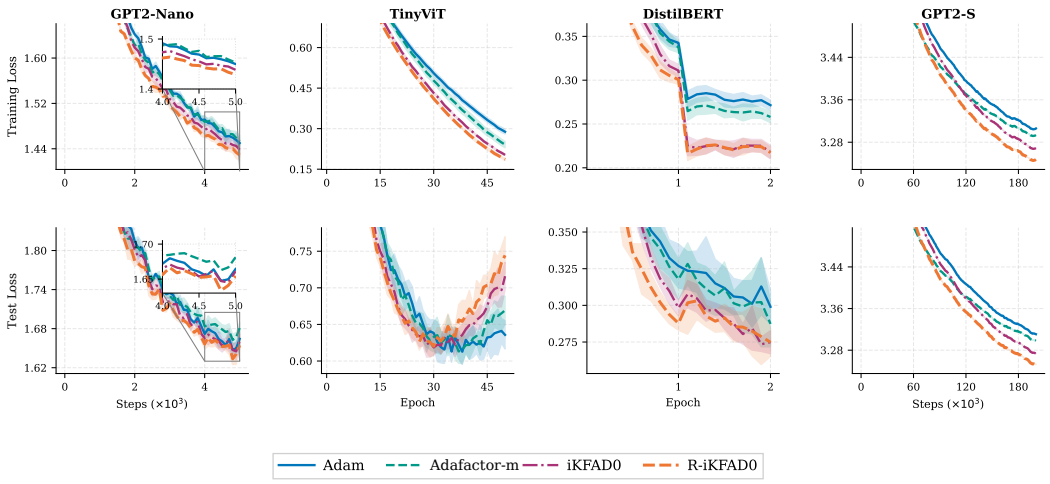

Figure 3: Loss curves for all four benchmarks, averaged over 10 random seeds (GPT2-S: a single seed). R-iKFAD0 matches or improves upon iKFAD0 on all tasks, while using approximately half the optimizer memory. Both methods are competitive with Adam and Adafactor-m, which require comparable or greater memory. Shaded regions denote one standard deviation.

Karoni et al. (2026) already established that iKFAD matches Adam on standard benchmarks, and Table 2 reproduces that result here. iKFAD0 equals or improves on Adam on three of the four benchmarks, the exception being TinyViT, by about one pooled standard deviation. Inspecting the loss curves for TinyViT, both iKFAD0 and R-iKFAD0 reach their minimum faster than Adam, which later attains a lower loss.

Moreover, R-iKFAD0 matches iKFAD0 on all four benchmarks while halving the auxiliary optimizer state (Table 5). Taken together, R-iKFAD0 performs like Adam at roughly half Adam's optimizer memory. These differences should be interpreted in light of the variability across seeds. On the three benchmarks with seed replication, iKFAD0 and R-iKFAD0 differ by less than one pooled standard deviation, whereas on TinyViT they agree to three decimal places. Where the friction methods do separate from Adam, the difference depends on the particular dataset: on DistilBERT they lead by about five pooled standard deviations, and on GPT2-S by 0.06 nats at a single seed, whereas on the one vision task Adam leads by about one. Given how small these differences are, we do not suggest that R-iKFAD0 outperforms Adam. Rather, the statement is that the friction tensor can be rank-1 factored without compromising performance.

Table 2: Test loss at the checkpoint selected by lowest *validation* loss (mean±std over 10 seeds, GPT2-S: single seed due to computational limitations). The test set was never used for model or hyper-parameter selection. Lower is better. R-iKFAD0 matches or improves upon iKFAD0 across all benchmarks while using half the optimizer memory.

Dataset	Adam	Adafactor-m	iKFAD0	R-iKFAD0 (ours)
GPT2-Nano	1.643 ± 0.008	1.661 ± 0.007	1.640 ± 0.008	$\mathbf{1.633 \pm 0.011}$
TinyViT	$\mathbf{0.593 \pm 0.015}$	0.596 ± 0.010	0.606 ± 0.011	0.606 ± 0.012
DistilBERT	0.292 ± 0.005	0.280 ± 0.007	0.269 ± 0.005	$\mathbf{0.268 \pm 0.003}$
GPT2-S	3.301	3.287	3.263	**3.242**

4.2 Ablation: $\gamma = 0$ versus $\gamma > 0$

Figure 4 compares R-iKFAD ($\gamma > 0$, with linear damping) against R-iKFAD0 ($\gamma = 0$, without linear damping) on GPT2-Nano, TinyViT, and DistilBERT (SST-2). Adding the γ term gave no consistent improvement over $\gamma = 0$ on these benchmarks under the tested procedure. This is consistent with the finding in Karoni et al. (2026) that the adaptive friction mechanism alone provides sufficient damping. We therefore recommend **R-iKFAD0** ($\gamma = 0$) as the practical default since it has one

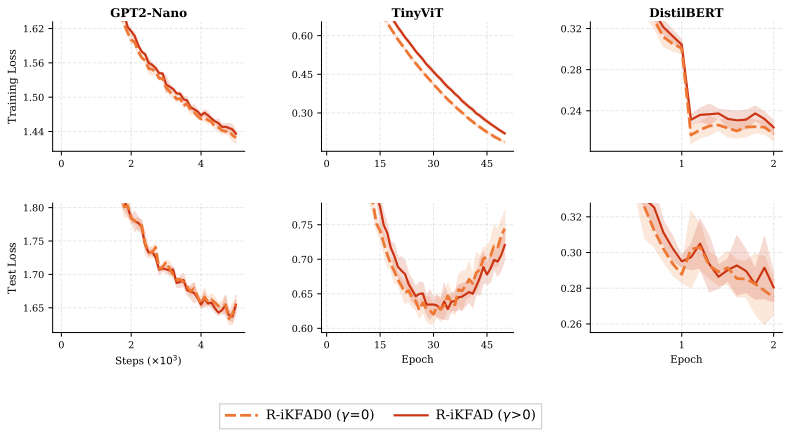

Figure 4: Training and test loss for R-iKFAD ($\gamma > 0$) vs. R-iKFAD0 ($\gamma = 0$) on GPT2-Nano, TinyViT, and DistilBERT (SST-2), averaged over 10 seeds (shaded = std). The two variants are indistinguishable in performance, supporting R-iKFAD0 as the recommended default.

fewer hyper-parameter. For the $\gamma > 0$ variant, we used the same h, α, and μ settings as R-iKFAD0 given in Appendix B and additionally swept γ log-uniformly over $[10^{-6}, 10]$.

4.3 Robustness to Hyperparameters

Figure 5 shows γ-h grid sweeps for R-iKFAD and iKFAD. Both methods exhibit broad regions of the tested grid with similar performance. In the R-iKFAD equations, an increase in momentum increases the factors R and C, which increases $\tilde{\xi}$ and hence the damping. In these sweeps, the rank-1 approximation did not substantially reduce the range of (γ, h) values giving low loss. The breadth of the plateau is a consequence of the feedback being negative: a step size large enough to inflate the momentum also inflates R and C, and thus $\tilde{\xi}$, so the damping increases proportionally. The same mechanism explains why γ adds little in the ablation study of Section 4.2: a constant damping term acts along a direction which the adaptive term already covers, but without responding to the change of state.

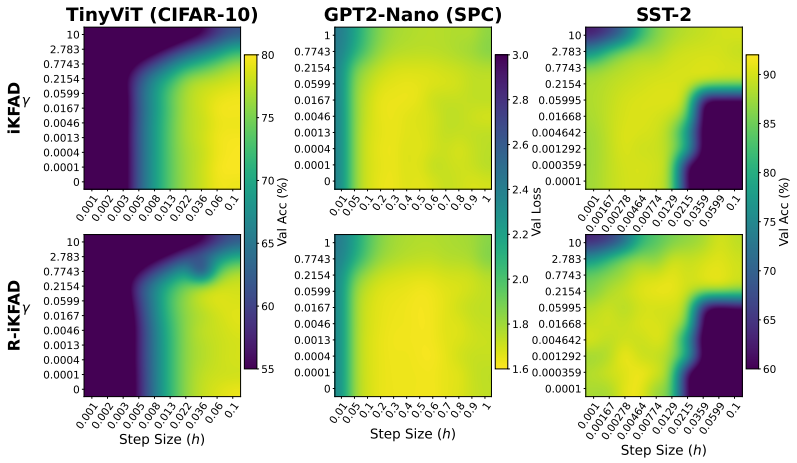

Figure 5: Two-dimensional γ–h sensitivity for iKFAD (top) and R-iKFAD (bottom). Each cell shows best performance at a fixed (h, γ) grid point. Brighter = better. R-iKFAD retains broad high-performance plateaus.

4.4 How Close is $\tilde{\xi}$ to the Best Rank-1 Friction?

Factorization could give a poor approximation for two reasons: ξ (iKFAD) may not be close to rank-1, or the construction of $\tilde{\xi}$ (R-iKFAD) may miss a good rank-1 approximation. By the Eckart-Young theorem, the smallest relative error of any rank-1 approximation is $\sqrt{1-\sigma_1^2/\sum_i \sigma_i^2}$.

Figure 6 compares this minimum with the error of $\tilde{\xi}$. We run iKFAD0 and update R-iKFAD's factors (R_ℓ, C_ℓ) alongside it, without using them in the optimizer updates. This lets us compare ξ and $\tilde{\xi}$ at the same iterates. Across the four benchmarks, the error of $\tilde{\xi}$ is between 0.07 and 0.85 percentage points above the minimum. On GPT2-S, the median gap across the 74 two-dimensional layers is 0.06 percentage points, and 61 layers are within one percentage point of the minimum. Relative to the minimum error, the gap is largest where ξ is nearly rank-1 and the minimum is close to zero.

The stable rank $\|\xi\|_F^2/\|\xi\|_2^2$ ranges from 1.18 on TinyViT to 2.01 on GPT2-Nano. Within GPT2-S, it varies by module: early attention value projections have stable ranks close to one, while query and key projections in the middle of the network have the largest stable ranks. These results suggest that most of the approximation error comes from the rank-1 restriction itself. The construction of $\tilde{\xi}$ adds little further error.

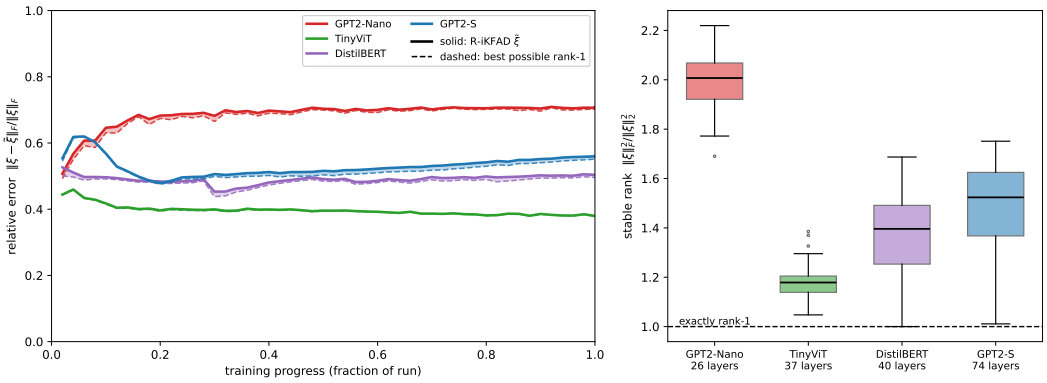

Figure 6: R-iKFAD's rank-1 friction against the best rank-1 approximation of the same tensor. **Left:** layer-averaged relative error $\|\xi - \tilde{\xi}\|_F/\|\xi\|_F$ (solid) against the Eckart–Young floor $\sqrt{1-\sigma_1^2/\sum_i \sigma_i^2}$ (dashed). The shaded region is the excess attributable to the construction, and is never wider than 0.85 pp. The horizontal axis is the fraction of each run completed, as the four benchmarks used different step budgets. **Right:** distribution over layers of the stable rank of ξ at the final step, where 1 would be exactly rank-1. Measurement is taken along iKFAD0 trajectories carrying R-iKFAD's factors as shadow states.

5 Conclusion

We introduced **R-iKFAD**, which replaces iKFAD's full friction tensor with a rank-1 outer product of momentum row and column sums, reducing per-layer friction state from $\mathcal{O}(mn)$ to $\mathcal{O}(m+n)$ while retaining a full momentum buffer. Across GPT2-Nano, TinyViT, DistilBERT and GPT2-S, R-iKFAD matches iKFAD in convergence speed, final performance and hyperparameter robustness at roughly half the auxiliary optimizer memory. The accompanying analysis, unlike the usual treatment of momentum-based dynamics, does not require positive linear damping. Under strong convexity we prove exponential convergence when $\gamma > 0$ and convergence to the minimizer when $\gamma = 0$. In the undamped case we demonstrate matching upper and lower bounds on the energy, of order t^{-1} when $\epsilon_{\text{stab}} = 0$ and of order $t^{-1/2}$ when $\epsilon_{\text{stab}} > 0$. Numerical integration of the dynamics agrees with both exponents. The analysis is continuous-time and deterministic, and assumes strong convexity. Extending it to the discrete-time, stochastic-gradient setting remains open.

AI USE STATEMENT

We used generative AI for the following purposes: We formulated the theorems ourselves, in particular the key results on algebraic rates, but generative AI was used to verify the proof including performing the numerical test of the algebraic rates and supplementing some intermediate steps in the proof. We also used generative AI to write the bash scripts that launched experiments on GPU and the visualization code behind the figures. We also used generative AI to improve the wording of the manuscript and to help draft some sentences, which the authors then revised. All scripts were checked carefully by the authors. We take full responsibility for the content of this work.

REPRODUCIBILITY STATEMENT

Anonymized Python source code for all four benchmarks is included in the supplementary material as a `.zip`. It contains the iKFAD and R-iKFAD implementations, training scripts, and hyperparameter sweep scripts, as well as per-benchmark environment specifications including seed numbers. Visualization code for Figures 3, 4 and 7 is included in a notebook. `theory_numerics/` contains the MATLAB code behind the numerical verification of the rates in Figure 2. Appendix B documents the hyperparameter sweep and training configurations.

REFERENCES

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2019.

Hédy Attouch, Zaki Chbani, Juan Peypouquet, and Patrick Redont. Fast convergence of inertial dynamics and algorithms with asymptotic vanishing viscosity. *Mathematical Programming*, 168 (1–2):123–175, 2018. doi: 10.1007/s10107-016-0992-8.

André Belotto Da Silva and Maxime Gazeau. A general system of differential equations to model first-order adaptive algorithms. *The Journal of Machine Learning Research*, 21(1):5072–5113, 2020.

Yusu Hong and Junhong Lin. Theoretical investigation of Adafactor for non-convex smooth optimization. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL `https://openreview.net/forum?id=QIv5aXEAcc`.

Aikaterini Karoni. *Higher-order damping mechanisms with applications in optimisation and machine learning*. PhD thesis, The University of Edinburgh, 2024.

Aikaterini Karoni, Benedict Leimkuhler, and Gabriel Stoltz. Friction-adaptive descent: A family of dynamics-based optimization methods. *Journal of Computational Dynamics*, 2023. doi: 10.3934/jcd.2023007.

Aikaterini Karoni, Rajit Rajpal, Benedict J. Leimkuhler, and Gabriel Stoltz. Adaptive momentum and nonlinear damping for neural network training. In *Forty-third International Conference on Machine Learning*, 2026. URL `https://openreview.net/forum?id=JWWpV4StVf`.

Andrej Karpathy. nanoGPT. `https://github.com/karpathy/nanoGPT`, 2022.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015*, 2015.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. Technical report, 2009.

Benedict Leimkuhler and Sebastian Reich. *Simulating Hamiltonian Dynamics*. Cambridge Monographs on Applied and Computational Mathematics. Cambridge University Press, 2005.

Son Nguyen, Lizhang Chen, Bo Liu, and Qiang Liu. Memory-efficient optimization with factorized Hamiltonian descent. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*. PMLR, 2025. URL https://proceedings.mlr.press/v258/nguyen25e.html.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter, 2020. URL https://arxiv.org/abs/1910.01108.

Noam Shazeer and Mitchell Stern. Adafactor: Adaptive learning rates with sublinear memory cost. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, 2018.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 2013.

Weijie Su, Stephen Boyd, and Emmanuel J. Candès. A differential equation for modeling Nesterov's accelerated gradient method: Theory and insights. *Journal of Machine Learning Research*, 17 (153):1–43, 2016.

Kan Wu, Jinnian Zhang, Houwen Peng, Mengchen Liu, Bin Xiao, Jianlong Fu, and Lu Yuan. TinyViT: Fast pretraining distillation for small vision transformers. In *European Conference on Computer Vision (ECCV)*, 2022.

A Update Rules for Splitting Discretisation

This appendix gives the explicit sub-step updates for the BACD splitting introduced in Section 2, and the treatment of the one sub-flow that does not admit a closed-form joint solution.

Table 3: Analytical sub-step update rules for the BACD splitting sequence over time step h.

Sub-step	Update Equation
B	$P^{(B)} = P_n - h\,\nabla f(X_n)$
A	$X_{n+1} = X_n + h\,P^{(B)}$
C	$\tilde{\xi}_n = \dfrac{R_n C_n^\top}{\mathbf{1}_m^\top R_n + \epsilon_{\text{stab}}}$
	$P^{(1)} = P^{(B)} \odot \exp\!\Big(-\frac{h}{2}\,\tilde{\xi}_n\Big)$
	$R_{n+1} = \mathrm{e}^{-\alpha h} R_n + \dfrac{1-\mathrm{e}^{-\alpha h}}{\mu\alpha}[P^{(1)}]^2 \mathbf{1}_n$
	$C_{n+1} = \mathrm{e}^{-\alpha h} C_n + \dfrac{1-\mathrm{e}^{-\alpha h}}{\mu\alpha}([P^{(1)}]^2)^\top \mathbf{1}_m$
	$\tilde{\xi}_{n+1} = \dfrac{R_{n+1} C_{n+1}^\top}{\mathbf{1}_m^\top R_{n+1} + \epsilon_{\text{stab}}}$
	$P^{(2)} = P^{(1)} \odot \exp\!\Big(-\frac{h}{2}\,\tilde{\xi}_{n+1}\Big)$
D	$P_{n+1} = P^{(2)} \odot \mathrm{e}^{-\gamma h}$ (*identity when* $\gamma = 0$)

Within sub-flow C, no closed-form joint solution exists for coupled P and (R, C). We apply a symmetric sequence: evaluating $\exp(-\frac{h}{2}\tilde{\xi})$ twice around an exact exponential ODE factor update improves integration accuracy within sub-flow C alone. Using the intermediate momentum $P^{(1)}$ for factor updates reduces forcing magnitude.

B Optimal Hyperparameters

Across models and tasks, the best-performing scales for μ (iKFAD0, R-iKFAD0) varied by several orders of magnitude. This variation is likely related to the scale of the momentum variables during training, since μ controls the strength of the adaptive friction terms. Normalizing the momentum variables across iterations could potentially reduce this sensitivity, but we leave this for future work.

Hyperparameters were selected by minimum validation loss during the Optuna sweeps. All reported results are computed on held-out test data.

Experimental configurations:

- **TinyViT – CIFAR-10**: batch size 128, 25 epochs, 100 trials. $h \in [10^{-6}, 10^{-1}]$, $\alpha \in [10^{-3}, 1]$, $\mu \in [10^{-8}, 10]$ for iKFAD0 and R-iKFAD0.
- **DistilBERT – SST-2**: batch size 16, 2 epochs, 80 trials. $h \in [10^{-6}, 10^{-1}]$, $\alpha \in [10^{-3}, 1]$, $\mu \in [10^{-8}, 10]$ for iKFAD0 and R-iKFAD0.
- **GPT2-Nano – Shakespeare**: batch size 16, 5000 steps for the hyperparameter sweeps and final reported runs, 500 trials. $h \in [10^{-6}, 5 \times 10^{-1}]$, $\alpha \in [10^{-5}, 10]$, $\mu \in [10^{-8}, 10]$ for iKFAD0 and R-iKFAD0.
- **GPT2-S – OpenWebText**: batch size 16, 75001 steps for the hyperparameter sweeps and 200001 steps for the final reported runs, 80 trials. $h \in [10^{-6}, 10^{-1}]$, $\alpha \in [10^{-3}, 1]$, $\mu \in [10^{-8}, 10]$ for iKFAD0 and $\mu \in [10^{-9}, 10^{-2}]$ for R-iKFAD0.

Table 4: Optimized hyperparameters for all experiments. All iKFAD0 and R-iKFAD0 entries have $\gamma = 0$. Dashes indicate parameters not used by the method.

	h	α	μ	β_1	β_2
TinyViT (CIFAR-10)					
iKFAD0	0.0725	0.0899	1.2643×10^{-5}	–	–
R-iKFAD0	0.0860	0.5327	9.4765×10^{-7}	–	–
Adam	5.5311×10^{-4}	–	–	0.8608	0.8852
Adafactor-m	5.6050×10^{-4}	–	–	0.8907	–
DistilBERT (SST-2)					
iKFAD0	0.0165	0.0574	9.8948×10^{-6}	–	–
R-iKFAD0	0.0423	0.0060	8.1874×10^{-7}	–	–
Adam	3.82×10^{-5}	–	–	0.9187	0.9825
Adafactor-m	2.5638×10^{-5}	–	–	0.9622	–
GPT2-Nano (Shakespeare)					
iKFAD0	0.4941	2.1955	4.5521×10^{-6}	–	–
R-iKFAD0	0.4842	2.8475	1.9407×10^{-6}	–	–
Adam	1.6803×10^{-3}	–	–	0.8876	0.9265
Adafactor-m	2.9923×10^{-3}	–	–	0.8893	–
GPT2-S (OpenWebText)					
iKFAD0	0.0996	0.0476	1.0429×10^{-5}	–	–
R-iKFAD0	0.0918	0.0579	1.3637×10^{-6}	–	–
Adam	6.4723×10^{-4}	–	–	0.8945	0.9945
Adafactor-m	7.2777×10^{-4}	–	–	0.8865	–

For all experiments, Adam was swept over $h \in [10^{-6}, 10^{-2}]$, $\beta_1 \in [0.85, 0.999]$, $\beta_2 \in [0.85, 0.999]$. Adafactor-m was swept over $h \in [10^{-5}, 10^{-2}]$ and $\beta_1 \in [0.85, 0.999]$, with no relative step or warmup. Adafactor-m has no tunable β_2: as originally proposed (Shazeer & Stern, 2018) and in standard practice, the second-moment decay follows the prescribed schedule $\beta_{2,t} = 1 - t^{-0.8}$

(`decay_rate` $= -0.8$), which is why its β_2 entries are left blank. We enable $\beta_1 > 0$ rather than the conventional momentum-free setting so that Adafactor-m occupies the same auxiliary-memory regime as R-iKFAD (Table 5). All sweeps used Optuna's (Akiba et al., 2019) TPE sampler. Table 4 reports the best hyperparameters found. iKFAD0 and R-iKFAD0 are the $\gamma = 0$ variants. All such entries satisfy $\gamma = 0$ by definition. For iKFAD0 and R-iKFAD0, h denotes the step size and μ the adaptive damping coefficient. For Adam and Adafactor-m, h is the learning rate.

C Memory Analysis

Table 5: Optimizer auxiliary state (MiB, float32), computed over all model parameters. R-iKFAD achieves approximately $2\times$ savings over both iKFAD and Adam because the rank-1 friction factors (R_ℓ, C_ℓ) are negligible relative to the full momentum buffer. The Adafactor-m baseline used throughout enables its optional first moment ($\beta_1 > 0$, see Section 4) and therefore carries an auxiliary footprint identical to that of R-iKFAD, making the two directly comparable at matched memory. Momentum-free Adafactor is included for reference.

Model	N	Adam	Adafactor-m	Adafactor	iKFAD	R-iKFAD
TinyViT	4.0M	30.7	15.6	0.2	30.7	15.6 (2.0×)
DistilBERT	67.0M	510.8	256.1	0.7	510.8	256.1 (2.0×)
GPT2-Nano	0.8M	6.1	3.1	< 0.1	6.1	3.1 (2.0×)
GPT2-S	124.4M	948.9	475.4	0.9	948.9	475.4 (2.0×)

D Accuracy Curves

Figure 7 gives the accuracy curves corresponding to the loss curves of Figure 3, for the two classification benchmarks. They are reported here rather than in Section 4 because they follow the loss curves closely.

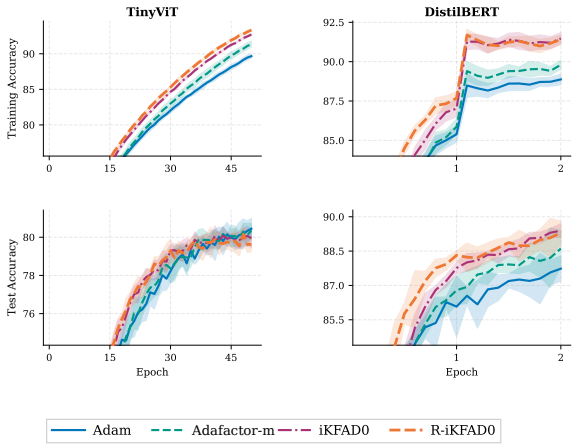

Figure 7: Accuracy averaged over 10 seeds. R-iKFAD0 matches iKFAD0 and Adam. Error bars denote one std.

E Prefactor Dependence

The prefactors depend on α, μ and ϵ_{stab} as the calculation of Section 3.1 predicts. If the asymptotics are $\mathcal{E} \sim c_1 \alpha\mu/t$ and $\mathcal{E} \sim c_2 \alpha\mu\sqrt{\epsilon_{\text{stab}}/2t}$, then the ratios in Table 6 are approximately constant. They were evaluated at $t = 10^4$ with $m = 3$, $n = 2$. The crossover row indicates the last downward crossing of $\mathbf{1}^\top R = \epsilon_{\text{stab}}$.

Table 6: Parameter dependence of the prefactors, evaluated at $t = 10^4$ with $m = 3$, $n = 2$. In the first row α and μ each range over a factor of 4, so $\alpha\mu$ ranges over a factor of 16, yet the ratio moves by less than 30%.

quantity	parameters varied	mean	spread (max/min)
$t\mathcal{E}/(\alpha\mu)$, $\epsilon_{\mathrm{stab}} = 0$	$\alpha, \mu \in \{\frac{1}{2}, 1, 2\}$	4.69	1.29
$\mathcal{E}\sqrt{t}/\big(\alpha\mu\sqrt{\epsilon_{\mathrm{stab}}/2}\big)$, $\epsilon_{\mathrm{stab}} > 0$	$\alpha, \mu \in \{\frac{1}{2}, 1\}$, $\epsilon_{\mathrm{stab}} \in \{0.05, 0.2\}$	2.31	1.14
$\mathcal{E}/(\alpha\mu\,\epsilon_{\mathrm{stab}})$ at the crossing	$\alpha, \mu \in \{\frac{1}{2}, 1\}$, $\epsilon_{\mathrm{stab}} \in \{0.05, 0.2\}$	0.87	1.21

The third row places the transition between the two regimes at $\mathcal{E} \approx 0.87\,\alpha\mu\,\epsilon_{\mathrm{stab}}$, confirming the crossover scale. For the scalar system $m = n = 1$ with $\alpha = \mu = 1$, evaluating at $t = 10^5$ gives $t\mathcal{E}/(\alpha\mu) = 0.909$ and $\mathcal{E}\sqrt{t}/(\alpha\mu\sqrt{\epsilon_{\mathrm{stab}}/2}) = 0.884$, both within about 10% of the value 1 the averaging calculation gives. For $m, n > 1$ the same relations hold with a dimension-dependent coefficient.

F Numerical Setup for Section 3.2

The initial entries of X and P were independent normal variables with standard deviations 0.4 and 0.3, and $R(0) = C(0) = 0$ throughout. The parameter sweeps used MATLAB R2026a, `rng(0)`, and `ode89` with relative tolerance 10^{-10} and absolute tolerance 10^{-12}. When $\epsilon_{\mathrm{stab}} = 0$ and $\mathbf{1}^\top R = 0$ the code set $\tilde{\xi} = 0$, and used (15) once $\mathbf{1}^\top R > 0$. The curves in Figure 2 were computed by BACD splitting with $m = 4$, $n = 3$, $\alpha = \mu = 1$, $h = 0.02$, final time 10^7 and random seed 1. The additional runs behind Table 1 used dimensions 2×2 and 6×5 with the same A, $X(0)$ and $P(0)$.

As consistency checks, for $m = 3$, $n = 2$ and $\alpha = \mu = 1$, the largest absolute difference between the `ode89` and BACD slopes was 0.018 over the three fitted intervals. Repeating the six BACD fits at $h = 0.01$ instead of $h = 0.02$ changed each slope by less than 8.9×10^{-4}.

G Rank-1 iKFAD Convergence Analysis

We analyze the continuous-time Rank-1 iKFAD dynamics for a matrix parameter $X \in \mathbb{R}^{m\times n}$ (the vector case follows by flattening). The system is given by

$$\dot{X} = P, \tag{19a}$$

$$\dot{P} = -\nabla f(X) - \tilde{\xi} \odot P - \gamma P, \tag{19b}$$

$$\dot{R} = \frac{1}{\mu}[P]^2\mathbf{1}_n - \alpha R, \tag{19c}$$

$$\dot{C} = \frac{1}{\mu}([P]^2)^\top\mathbf{1}_m - \alpha C, \tag{19d}$$

where $P \in \mathbb{R}^{m\times n}$, $R \in \mathbb{R}^m$, $C \in \mathbb{R}^n$, and $\gamma \geq 0$, $\alpha, \mu > 0$ are hyperparameters. Boundedness (Lemma 4) holds for $\gamma \geq 0$. The exponential convergence results below additionally require $\gamma > 0$. The rank-1 friction tensor $\tilde{\xi} \in \mathbb{R}^{m\times n}$ is defined element-wise as

$$\tilde{\xi}_{ij} = \frac{R_i C_j}{\sum_{k=1}^m R_k + \epsilon_{\mathrm{stab}}},$$

with $\epsilon_{\mathrm{stab}} > 0$ a small constant preventing division by zero. The notation $\odot$ denotes element-wise multiplication, $[\,\cdot\,]^2$ element-wise squaring, and $\langle\cdot,\cdot\rangle_F$ the Frobenius inner product.

G.1 Dynamical properties and boundedness

Throughout this appendix X^* denotes a global minimizer of f. Under the hypotheses used below such a point exists, because f is continuous and coercive, and it satisfies

$$\nabla f(X^*) = 0 \qquad \text{and} \qquad f(X) \geq f(X^*) \quad \text{for all } X. \tag{20}$$

Lemma 4 (Boundedness of solutions)**.** *Assume $f \in C^2$ (so ∇f is automatically locally Lipschitz), $f(X) \to +\infty$ as $\|X\|_F \to +\infty$, and let X^* be a global minimizer of f as in* (20)*. Then for any initial condition (X_0, P_0, R_0, C_0) with $R_0, C_0 \geq 0$ entrywise, the solution of* (19a)–(19d) *is well-defined for all $t \geq 0$, and there exists a compact set $\mathcal{K}$ containing the trajectory. In particular, with*

$$P_{\max}^2 := 2\big(f(X_0) - f(X^*) + \tfrac{1}{2}\|P_0\|_F^2\big), \quad R_{\max} := \|R_0\| + \frac{P_{\max}^2}{\alpha\mu}, \quad C_{\max} := \|C_0\| + \frac{P_{\max}^2}{\alpha\mu},$$

all of which are finite and independent of t, we have for all $t \geq 0$

$$0 \leq R_i(t) \leq R_{\max}, \quad 0 \leq C_j(t) \leq C_{\max}, \quad \|P(t)\|_F \leq P_{\max},$$

and moreover the friction tensor obeys the bound

$$0 \leq \tilde{\xi}_{ij}(t) \leq \xi_{\max} := \min\Big\{C_{\max},\ \frac{R_{\max}C_{\max}}{\epsilon_{\mathrm{stab}}}\Big\}. \tag{21}$$

Note that $P_{\max}$ depends on neither μ nor γ.

Proof. The ODEs (19c)–(19d) have non-negative forcing and linear decay, so $R_0, C_0 \geq 0$ implies $R_i(t), C_j(t) \geq 0$ for all $t \geq 0$, and hence $\tilde{\xi}_{ij} \geq 0$ throughout.

Consider the reduced energy

$$\mathcal{E}(X, P) = f(X) - f(X^*) + \tfrac{1}{2}\|P\|_F^2,$$

which omits the R, C contributions. Differentiating along trajectories and cancelling the gradient terms,

$$\dot{\mathcal{E}} = \langle \nabla f(X), P\rangle_F + \langle P, -\nabla f(X) - \tilde{\xi} \odot P - \gamma P\rangle_F = -\gamma\|P\|_F^2 - \sum_{i,j} \tilde{\xi}_{ij} P_{ij}^2 \leq 0, \tag{22}$$

both terms being non-positive because $\gamma \geq 0$ and $\tilde{\xi} \geq 0$ entrywise. No bound on $\|P\|_F$ is used in deriving (22), so the estimate is not circular. Consequently $\mathcal{E}(t) \leq \mathcal{E}(0)$ for all t, which yields at once

$$\|P(t)\|_F^2 \leq 2\mathcal{E}(0) =: P_{\max}^2,$$

and, since $f(X) - f(X^*) \leq \mathcal{E}(0)$ with f coercive, confinement of $X(t)$ to a compact sublevel set of f.

With $\|P\|_F \leq P_{\max}$ now established independently, the factors follow from the explicit linear filter. Integrating (19c),

$$R(t) = \mathrm{e}^{-\alpha t}R_0 + \frac{1}{\mu}\int_0^t \mathrm{e}^{-\alpha(t-s)}[P(s)]^2\mathbf{1}_n \,\mathrm{d}s,$$

and using the sharp, dimension-free estimate $\big\|[P]^2\mathbf{1}_n\big\| \leq \|P\|_F^2 \leq P_{\max}^2$ (the entries of $[P]^2\mathbf{1}_n$ are non-negative and sum to $\|P\|_F^2$, so its 2-norm is at most its 1-norm),

$$\|R(t)\| \leq \|R_0\| + \frac{P_{\max}^2}{\alpha\mu} =: R_{\max},$$

with the analogous bound $C_{\max} := \|C_0\| + P_{\max}^2/(\alpha\mu)$. All four variables therefore remain in a compact set $\mathcal{K}$ for all $t \geq 0$, and global existence follows from the standard continuation criterion for locally Lipschitz vector fields.

Finally we establish (21). Two bounds are possible. Bounding the denominator below by ϵ_{stab} gives $\tilde{\xi}_{ij} \leq R_{\max}C_{\max}/\epsilon_{\mathrm{stab}}$, which diverges as $\epsilon_{\mathrm{stab}} \downarrow 0$. The second is sharper and uniform in ϵ_{stab}: since every $R_k \geq 0$ we have $R_i \leq \mathbf{1}_m^\top R$, hence

$$\tilde{\xi}_{ij} = \frac{R_i}{\mathbf{1}_m^\top R + \epsilon_{\mathrm{stab}}}\, C_j \leq C_j \leq C_{\max},$$

and (21) is the minimum of the two. The constant $\epsilon_{\mathrm{stab}} > 0$ keeps the friction well defined at $R = 0$, which is how the implementation initializes the factors. The case $\epsilon_{\mathrm{stab}} = 0$ is treated separately in Lemma 5. The bound (21) is independent of ϵ_{stab}, so the stabilizer does not affect the exponential rate of Theorem 1. At $\gamma = 0$, however, it does determine the algebraic exponent (Theorem 3). □

G.2 GEOMETRIC CONVERGENCE

Theorem 1. Suppose that f is m_f-strongly convex and has an L_f-Lipschitz gradient. Let $\gamma, \alpha, \mu, \epsilon_{\mathrm{stab}} > 0$, and let $R(0)$ and $C(0)$ be componentwise nonnegative. Then there are constants $M, \rho > 0$ such that

$$f(X(t)) - f(X^*) + \|X(t) - X^*\|_F^2 + \|P(t)\|_F^2 + \|R(t)\|_2^2 + \|C(t)\|_2^2 \le Me^{-\rho t} \tag{23}$$

for every $t \ge 0$.

Proof of Theorem 1. Put $Y = X - X^*$ and $d_{\max} = \gamma + \xi_{\max}$. Choose

$$0 < \delta \le \min\left\{ \frac{\sqrt{m_f}}{2}, \frac{\gamma}{2\left(1 + d_{\max}^2/(2m_f)\right)} \right\}. \tag{24}$$

Here $\xi_{\max}$ is the bound in Lemma 4. Define

$$V = f(X) - f(X^*) + \frac{1}{2}\|P\|_F^2 + \delta\langle Y, P\rangle_F.$$

Young's inequality gives

$$V \ge \frac{m_f}{4}\|Y\|_F^2 + \frac{1}{4}\|P\|_F^2. \tag{25}$$

It also gives

$$V \le \left(\frac{L_f}{2} + \frac{m_f}{4}\right)\|Y\|_F^2 + \frac{3}{4}\|P\|_F^2. \tag{26}$$

Differentiating along the equations gives

$$\begin{aligned} \dot{V} &\le -(\gamma - \delta)\|P\|_F^2 - \delta m_f \|Y\|_F^2 + \delta d_{\max}\|Y\|_F\|P\|_F \\ &\le -\frac{\delta m_f}{2}\|Y\|_F^2 - \frac{\gamma}{2}\|P\|_F^2. \end{aligned} \tag{27}$$

The last line follows from Young's inequality and (24).

Set

$$\rho_0 := \min\left\{ \frac{2\delta m_f}{2L_f + m_f}, \frac{2\gamma}{3} \right\}.$$

Equations (26) and (27) imply

$$\dot{V} \le -\rho_0 V.$$

Therefore $V(t) \le e^{-\rho_0 t}V(0)$.

The lower bound (25) now gives exponential decay of $X - X^*$ and P. The variation-of-constants formulas for R and C then give their exponential decay. For example,

$$R_i(t) = e^{-\alpha t}R_i(0) + \frac{1}{\mu}\int_0^t e^{-\alpha(t-s)}\sum_{k=1}^{n} P_{ik}(s)^2 \, \mathrm{d}s.$$

The convolution on the right decays exponentially. The same argument applies to C. Taking, for example,

$$0 < \rho < \min\{\rho_0, 2\alpha\}$$

and increasing M if necessary proves (17). The same bounds prevent finite-time escape, so the solution is defined for every $t \ge 0$. □

The proof uses the fixed term $-\gamma\|P\|_F^2$ to control the position–momentum cross term. When $\gamma = 0$, this control is absent. The next subsection therefore treats this case separately.

G.3 THE CASE $\gamma = 0$

In our experiments we observed that $\gamma = 0$ gave optimal or near-optimal results. The next theorem proves convergence of the continuous system in this case.

Theorem 2. Suppose that f is strongly convex and has a locally Lipschitz gradient. Let $\alpha, \mu, \epsilon_{\mathrm{stab}} > 0$, and suppose that $R(0)$ and $C(0)$ are componentwise nonnegative. Then every solution of (19a)–(19d) with $\gamma = 0$ converges to

$$(X, P, R, C) = (X_\star, 0, 0, 0),$$

where $X_\star$ is the unique minimizer of f.

Proof of Theorem 2. Set

$$\mathcal{H}(X, P) = f(X) + \frac{1}{2}\|P\|_F^2.$$

Along a solution with $\gamma = 0$,

$$\begin{aligned}\dot{\mathcal{H}} &= \langle \nabla f(X), P\rangle_F + \langle P, -\nabla f(X) - \tilde{\xi} \odot P\rangle_F \\ &= -\sum_{i,j} \tilde{\xi}_{ij} P_{ij}^2 \le 0. \end{aligned} \tag{28}$$

Thus $\mathcal{H}$ is nonincreasing. Its sublevel sets are bounded by strong convexity. Lemma 4 gives global existence.

The factors satisfy

$$R_i(t) = e^{-\alpha t} R_i(0) + \frac{1}{\mu}\int_0^t e^{-\alpha(t-s)} \sum_{k=1}^{n} P_{ik}(s)^2 \, \mathrm{d}s, \tag{29}$$

$$C_j(t) = e^{-\alpha t} C_j(0) + \frac{1}{\mu}\int_0^t e^{-\alpha(t-s)} \sum_{\ell=1}^{m} P_{\ell j}(s)^2 \, \mathrm{d}s. \tag{30}$$

These formulas show that the factors remain nonnegative. Since P is bounded, they also show that R and C are bounded. The complete trajectory is therefore precompact.

It remains to identify the largest invariant subset of $\{\dot{\mathcal{H}} = 0\}$. Equation (28) and $\tilde{\xi}_{ij} \ge 0$ imply that

$$\tilde{\xi}_{ij} P_{ij}^2 = 0 \quad \text{for every } i, j$$

on this set. Suppose that $P_{ij} \neq 0$ at some point of a trajectory in the set. Continuity makes $P_{ij}^2 > 0$ on a time interval. Because $\epsilon_{\mathrm{stab}} > 0$, the displayed equality then implies $R_i C_j = 0$ throughout that interval.

This is impossible. If $R_i = 0$ at any time in the interval, then

$$\dot{R}_i = \frac{1}{\mu}\sum_k P_{ik}^2 > 0.$$

Similarly, if $C_j = 0$, then

$$\dot{C}_j = \frac{1}{\mu}\sum_\ell P_{\ell j}^2 > 0.$$

Moreover, the variation-of-constants formulas show that a nonnegative factor, once positive, stays positive. Hence both R_i and C_j become positive in the interval, contradicting $R_i C_j = 0$. We conclude that every trajectory in $\{\dot{\mathcal{H}} = 0\}$ has $P = 0$.

If $P = 0$ along an invariant trajectory, then $\nabla f(X) = 0$ and $X = X^*$. The largest invariant subset of $\{\dot{\mathcal{H}} = 0\}$ is therefore

$$\{(X^*, 0, R, C) : R, C \ge 0\}.$$

LaSalle's invariance principle gives

$$X(t) \longrightarrow X^*, \qquad P(t) \longrightarrow 0.$$

Finally, equations (29)–(30) and $P(t) \to 0$ give $R(t), C(t) \to 0$. □

G.3.1 ALGEBRAIC RATES

The formal calculation above identifies the exponents. The next result establishes them rigorously, with matching upper and lower bounds in both regimes.

Throughout this subsection $\gamma = 0$ in (19a)–(19d), and the friction is

$$\tilde{\xi}_{ij} = \frac{R_i C_j}{s + \epsilon_{\mathrm{stab}}}, \qquad s = \sum_i R_i. \tag{31}$$

When $\epsilon_{\mathrm{stab}} = 0$, set $\tilde{\xi} = 0$ at $s = 0$. Since $\dot{R}_i \geq -\alpha R_i$ and $\dot{C}_j \geq -\alpha C_j$, nonnegative initial factors remain nonnegative, so $s \geq 0$ along the flow. We write $\|\tilde{\xi}\|_\infty = \max_{i,j} |\tilde{\xi}_{ij}|$.

Let X^* be the minimizer of f, and define

$$\mathcal{E}(t) = f(X(t)) - f(X^*) + \frac{1}{2}\|P(t)\|_F^2. \tag{32}$$

Theorem 3. Suppose that f is strongly convex, its gradient is locally Lipschitz, and it is C^2 in a neighbourhood of X^*. Let $\gamma = 0$ and $\alpha, \mu > 0$, and let the initial factors be componentwise nonnegative.

If $\epsilon_{\mathrm{stab}} > 0$ and $\mathcal{E}(0) > 0$, there are $a, A > 0$ such that

$$\frac{a}{\sqrt{1+t}} \leq \mathcal{E}(t) \leq \frac{A}{\sqrt{1+t}}, \qquad t \geq 0.$$

If $\epsilon_{\mathrm{stab}} = 0$, assume in addition that $R(0) = C(0) = 0$, and use the convention following (31). If $\mathcal{E}(0) > 0$, there are $b, B > 0$ such that

$$\frac{b}{1+t} \leq \mathcal{E}(t) \leq \frac{B}{1+t}, \qquad t \geq 0.$$

The proof uses two estimates. A lower bound on the energy lost over a fixed time window gives the upper bound on $\mathcal{E}$. An upper bound on the instantaneous loss gives the lower bound on $\mathcal{E}$.

Lemma 5 (The zero-stabilizer phase space)**.** *Suppose that $\epsilon_{\mathrm{stab}} = 0$ and $\sum_i R_i = \sum_j C_j =: s$. The convention $\tilde{\xi} = 0$ at $s = 0$ defines a locally Lipschitz friction tensor on*

$$\mathcal{B} = \Big\{ R \geq 0,\ C \geq 0,\ \sum_i R_i = \sum_j C_j \Big\}.$$

Moreover,

$$0 \leq \tilde{\xi}_{ij} \leq C_j \leq s. \tag{33}$$

Proof. The bound follows from $R_i \leq s$. For local Lipschitz continuity, consider two points of $\mathcal{B}$ with totals $0 \leq s \leq s'$. If $s = s' = 0$, both friction tensors vanish. If $s = 0 < s'$, then

$$|\tilde{\xi}_{ij} - \tilde{\xi}'_{ij}| = \frac{R'_i C'_j}{s'} \leq s' = |s - s'|.$$

If $s > 0$, write

$$R = s\rho, \quad C = s\sigma, \qquad R' = s'\rho', \quad C' = s'\sigma',$$

where the four normalized vectors lie in the appropriate simplices. Then

$$\big|s\rho_i\sigma_j - s'\rho'_i\sigma'_j\big| \leq |s - s'| + s|\rho_i - \rho'_i| + s|\sigma_j - \sigma'_j|,$$

and

$$s|\rho_i - \rho'_i| \leq |R_i - R'_i| + |s - s'|, \qquad s|\sigma_j - \sigma'_j| \leq |C_j - C'_j| + |s - s'|.$$

Since $|s - s'| \leq \|R - R'\|_1$, these estimates give the required Lipschitz bound. □

Lemma 6 (Kinetic energy on a fixed window)**.** *Assume the hypotheses of Theorem 3, and suppose a solution of* (19a)–(19d) *satisfies*

$$X(t) \longrightarrow X^*, \qquad P(t) \longrightarrow 0, \qquad \|\tilde{\xi}(t)\|_\infty \longrightarrow 0.$$

Then for every $T > 0$ there are $c_0 > 0$ and $t_0 \geq 0$, both depending on the solution, such that

$$\int_t^{t+T} \|P(u)\|_F^2 \, du \geq c_0 \mathcal{E}(t) \qquad (t \geq t_0). \tag{34}$$

Proof. Fix $T > 0$. If the result is false, for every k there is $t_k \geq k$ violating (34) with $c_0 = 1/k$. The integral is nonnegative, so $e_k = \mathcal{E}(t_k) > 0$, and

$$\frac{1}{e_k} \int_{t_k}^{t_k+T} \|P(u)\|_F^2 \, du \longrightarrow 0. \tag{35}$$

Set, for $0 \leq u \leq T$,

$$y_k(u) = \frac{X(t_k+u) - X^*}{\sqrt{e_k}}, \qquad p_k(u) = \frac{P(t_k+u)}{\sqrt{e_k}}.$$

The assumed convergence gives $e_k \to 0$. Energy monotonicity and strong convexity make y_k and p_k uniformly bounded. On bounded sets of y, the C^2 assumption gives, uniformly,

$$\frac{\nabla f(X^* + \sqrt{e_k} y)}{\sqrt{e_k}} \longrightarrow Hy, \qquad H = \nabla^2 f(X^*).$$

The equations and the uniform convergence of $\tilde{\xi}(t_k + u)$ to zero on $[0, T]$ now give uniform bounds on $\dot{y}_k$ and $\dot{p}_k$. Hence there is a constant L, independent of k, such that

$$\|y_k(u) - y_k(v)\|_F + \|p_k(u) - p_k(v)\|_F \leq L|u - v| \qquad (0 \leq u, v \leq T).$$

The Arzelà–Ascoli theorem therefore gives a subsequence on which both sequences converge uniformly, say to y and p. For this subsequence, the equations are

$$\begin{aligned} y_k(u) &= y_k(0) + \int_0^u p_k(v) \, dv, \\ p_k(u) &= p_k(0) - \int_0^u \left[\frac{\nabla f(X^* + \sqrt{e_k} y_k(v))}{\sqrt{e_k}} + \tilde{\xi}(t_k + v) \odot p_k(v) \right] dv. \end{aligned}$$

Uniform convergence gives $p_k \to p$. The Taylor estimate above gives uniform convergence of the scaled gradient to Hy, while $\tilde{\xi}(t_k + v) \odot p_k(v) \to 0$ uniformly. Taking the limit in these identities gives

$$y(u) = y(0) + \int_0^u p(v) \, dv, \qquad p(u) = p(0) - \int_0^u Hy(v) \, dv.$$

Thus the limits solve

$$\dot{y} = p, \qquad \dot{p} = -Hy. \tag{36}$$

Taylor expansion of f at X^* gives

$$\frac{1}{2} \langle Hy(0), y(0) \rangle + \frac{1}{2} \|p(0)\|_F^2 = 1.$$

Equation (35) gives $\int_0^T \|p(u)\|_F^2 \, du = 0$. Thus $p = 0$ on $[0, T]$. Equation (36) gives $Hy = 0$. Strong convexity makes H positive definite, so $y = 0$, contradicting the normalized energy. □

Proof of Theorem 3. Boundedness and convergence. Differentiating (32) gives

$$\dot{\mathcal{E}}(t) = -\sum_{i,j} \tilde{\xi}_{ij}(t) P_{ij}(t)^2. \tag{37}$$

Thus $\mathcal{E}$ is nonincreasing. Strong convexity bounds X and P. Variation of constants gives

$$R_i(t) = e^{-\alpha t} R_i(0) + \frac{1}{\mu} \int_0^t e^{-\alpha(t-u)} \sum_j P_{ij}(u)^2 \, du, \tag{38}$$

$$C_j(t) = e^{-\alpha t} C_j(0) + \frac{1}{\mu} \int_0^t e^{-\alpha(t-u)} \sum_i P_{ij}(u)^2 \, du. \tag{39}$$

Hence the factors are also bounded. Local Lipschitz continuity, together with boundedness, gives global existence. In the zero-stabilizer case, local Lipschitz continuity is understood on the phase space in Lemma 5. If $\mathcal{E}(0) = 0$, then $X = X^*$, $P = 0$, and the energy remains zero. Assume from now on that $\mathcal{E}(0) > 0$.

Consider a complete bounded trajectory in the ω-limit set on which $\dot{\mathcal{E}} = 0$. Boundedness of the complete trajectory gives

$$R_i(t) = \frac{1}{\mu} \int_{-\infty}^t e^{-\alpha(t-u)} \sum_j P_{ij}(u)^2 \, du, \qquad C_j(t) = \frac{1}{\mu} \int_{-\infty}^t e^{-\alpha(t-u)} \sum_i P_{ij}(u)^2 \, du.$$

If some $P_{ij} \neq 0$, it remains nonzero on a short interval, and these formulas make both R_i and C_j positive on a nonempty subinterval. This makes $\tilde{\xi}_{ij} P_{ij}^2 > 0$, a contradiction. Hence $P = 0$ on every such trajectory. Invariance then gives $\nabla f(X) = 0$, and thus $X = X^*$. LaSalle's invariance principle and the filter formulas give

$$X(t) \to X^*, \qquad P(t) \to 0, \qquad R(t), C(t) \to 0. \tag{40}$$

For $\epsilon_{\text{stab}} = 0$, the two factor totals satisfy the same scalar equation. The initialization therefore keeps the solution in $\mathcal{B}$, and Lemma 5 applies. In both cases, (40) gives $\|\tilde{\xi}(t)\|_\infty \to 0$. Since the friction is also bounded, $\|\tilde{\xi}\|_\infty \leq M_0$ for a fixed M_0, so (37) gives $-\dot{\mathcal{E}} \leq M_0 K \leq 2M_0 \mathcal{E}$, where $K = \|P\|_F^2$. Hence $\mathcal{E}(t) \geq \mathcal{E}(0) e^{-2M_0 t} > 0$.

The factor totals. Let q denote either $s = \sum_i R_i$ or $\bar{s} = \sum_j C_j$. Both satisfy

$$\dot{q} = \frac{K}{\mu} - \alpha q.$$

Set $w = q/\mathcal{E}$. Since $K \leq 2\mathcal{E}$, equation (37) gives

$$\dot{w} \leq \frac{2}{\mu} - \alpha w + 2\|\tilde{\xi}\|_\infty w.$$

The last coefficient tends to zero. Scalar comparison therefore gives constants $Z_R, Z_C > 0$ such that, for all sufficiently large t,

$$s(t) \leq Z_R \mathcal{E}(t), \qquad \bar{s}(t) \leq Z_C \mathcal{E}(t). \tag{41}$$

Lemma 6 also applies. Choose t_1 so that (34) and (41) hold for $t \geq t_1$.

A common window estimate. Let $d = mn$. For each $t \geq t_1$, choose (i, j) such that

$$G := \int_t^{t+T} P_{ij}(u)^2 \, du \geq \frac{c_0}{d} \mathcal{E}(t). \tag{42}$$

For $t \leq u \leq t + T$, put

$$G(u) = \int_t^u P_{ij}(v)^2 \, dv,$$

so that $G(t) = 0$ and $G(t+T) = G$. Equations (38) and (39), with their nonnegative omitted terms, give

$$R_i(u) \geq \frac{e^{-\alpha T}}{\mu} G(u), \qquad C_j(u) \geq \frac{e^{-\alpha T}}{\mu} G(u).$$

It follows that

$$\int_t^{t+T} R_i(u)C_j(u)P_{ij}(u)^2\,du \geq \frac{e^{-2\alpha T}}{\mu^2}\int_t^{t+T} G(u)^2G'(u)\,du$$
$$= \frac{e^{-2\alpha T}}{3\mu^2}G^3 \geq c_1\mathcal{E}(t)^3, \tag{43}$$

where $c_1 > 0$ is fixed.

Upper bound when $\epsilon_{\text{stab}} > 0$. Boundedness of R gives $s(u) + \epsilon_{\text{stab}} \leq M$ for a fixed M. Integrating (37), retaining the single index chosen above, and using (43) yields

$$\mathcal{E}(t) - \mathcal{E}(t+T) \geq c_2\mathcal{E}(t)^3. \tag{44}$$

Upper bound when $\epsilon_{\text{stab}} = 0$. The solution is not the equilibrium, and (38) gives $s(t) > 0$ for every $t > 0$. Throughout a window $[t, t+T]$ with $t \geq t_1$, (41) and monotonicity give

$$s(u) \leq Z_R\mathcal{E}(u) \leq Z_R\mathcal{E}(t).$$

Equations (37) and (43) now give

$$\mathcal{E}(t) - \mathcal{E}(t+T) \geq c_3\mathcal{E}(t)^2. \tag{45}$$

Iteration of the window estimates. Set $e_k = \mathcal{E}(t_1 + kT)$. If some $e_k = 0$, then $\mathcal{E} \equiv 0$ and the bound is trivial, so assume $e_k > 0$ for every k. If

$$e_k - e_{k+1} \geq ce_k^q,$$

then the mean value theorem gives

$$e_{k+1}^{-(q-1)} - e_k^{-(q-1)} \geq (q-1)c.$$

Thus $e_k = O(k^{-1/(q-1)})$. Monotonicity of $\mathcal{E}$ extends the bound from the grid points to every $t \geq t_1$. Increasing the constant covers $[0, t_1]$. Taking $q = 3$ in (44) and $q = 2$ in (45) gives the two upper bounds.

Lower bound when $\epsilon_{\text{stab}} = 0$. By (33), (37), and (41),

$$-\dot{\mathcal{E}} \leq sK \leq 2Z_R\mathcal{E}^2, \qquad \text{that is,} \qquad \frac{d}{dt}\frac{1}{\mathcal{E}} \leq 2Z_R,$$

for $t \geq t_1$. Integrating from t_1 gives

$$\frac{1}{\mathcal{E}(t)} \leq \frac{1}{\mathcal{E}(t_1)} + 2Z_R(t - t_1), \qquad t \geq t_1,$$

and hence a constant $b_1 > 0$ with $\mathcal{E}(t) \geq b_1/(1+t)$ for $t \geq t_1$.

Lower bound when $\epsilon_{\text{stab}} > 0$. Since $R_i \leq s$, $C_j \leq \bar{s}$, and $K \leq 2\mathcal{E}$, (37) and (41) give

$$\tilde{\xi}_{ij} \leq \frac{s\bar{s}}{\epsilon_{\text{stab}}}, \qquad -\dot{\mathcal{E}} \leq \frac{s\bar{s}}{\epsilon_{\text{stab}}}K \leq \frac{2Z_RZ_C}{\epsilon_{\text{stab}}}\mathcal{E}^3, \qquad \text{that is,} \qquad \frac{d}{dt}\frac{1}{\mathcal{E}^2} \leq \frac{4Z_RZ_C}{\epsilon_{\text{stab}}},$$

for $t \geq t_1$. Integrating from t_1 gives

$$\frac{1}{\mathcal{E}(t)^2} \leq \frac{1}{\mathcal{E}(t_1)^2} + \frac{4Z_RZ_C}{\epsilon_{\text{stab}}}(t - t_1), \qquad t \geq t_1,$$

and hence a constant $a_1 > 0$ with $\mathcal{E}(t) \geq a_1/\sqrt{1+t}$ for $t \geq t_1$. Reducing a_1 and b_1, if necessary, extends both lower bounds to $[0, t_1]$. □